

Decodability Is Not Computation: Examining Probing Claims

Kabir Singh Grewal

University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
ska@unc.edu

Abstract

High probe accuracy does not by itself establish that a language model computes and uses the decoded value. We examine this inference in a syntax-sensitive arithmetic testbed: Llama-3.1-8B often fails multiplication when an operand is wrapped in parentheses or an additive identity. In the primary operand band, an input-only linear baseline nearly matches the original hidden-state probe’s regression $R^2 \approx 0.96$. A classical variance decomposition gives an analytical headroom of only 0.031 for this baseline class. In wider bands, selection-separated reanalysis instead finds positive activation advantages, which increase with regularization tuning; joint input-and-activation probes also improve over the baseline. These predictive gains are distinct from exact integer recovery and causal use. In the primary band, tested answer-site steering and state transplants produce small average output shifts and no observed donor-answer emissions, whereas operand-site interventions produce large effects. Operand access is necessary, with a strong bottleneck at one attention layer. The causal tests do not cover the wider-band predictive gains. We advocate explicit input baselines, matched causal tests, and positive controls, with conclusions restricted to the distributions, representations, and interventions actually examined.

1 Introduction

A probe predicts a target from a model’s hidden state. Its success establishes accessibility to that estimator, but an account of the model’s computation requires more: identifying how a quantity is produced and how it affects the answer (Belinkov, 2022). Input features can make a target easy to predict even when the model does not use the hypothesized intermediate. Conversely, an activation advantage over an input baseline still does not identify the causal route that uses it.

We study these distinctions in multiplication. For Llama-3.1-8B, changing $B * C = \text{to } (\emptyset +$

$B) * C =$ preserves the mathematical answer but substantially changes accuracy. A probe nevertheless predicts the product from the last prompt position, the *answer site*, with high regression R^2 . The behavioral controls show that this is a format-sensitive testbed: for the base model, parentheses alone reproduce much of the failure. We therefore use the contrast to examine probing claims, without treating it as an isolated test of general compositional ability.

Our evidence separates three questions. First, how much predictive performance is available from declared input features? In the primary band [20, 49], an operand-only baseline nearly matches the original fixed-regularization activation probe. We call the largest possible improvement over a baseline its *headroom*. For independent operands and an additive baseline, a classical variance decomposition computes that ceiling before model execution. This is an effect-size diagnostic, not a power calculation.

Second, do activations add predictive value? A corrected analysis of saved wider-band activations separates layer and penalty selection from evaluation and compares both standalone and concatenated probes. The answer is yes in these bands, especially after tuning regularization. Thus the narrow-band comparison cannot establish that all product predictability is an *operand echo*, meaning reconstruction from linearly accessible operands rather than a distinct product intermediate. Similar standalone scores are consistent with that account; they do not identify the activation probe’s features.

Third, can the decoded value control the answer through the tested site? In the primary band, probe-direction steering and answer-site state swaps produce no observed donor-product emissions. Matched key-value interventions at operand positions produce much larger effects. Blocking operand access collapses successful multiplication, particularly at layer 15. These interventions

constrain a sufficient, transplantable product-store interpretation of the tested answer-site channels. They do not exclude other subspaces, gating, re-computation, or causal effects in the wider bands.

The contribution is a practical audit of the inference from decodability to computation: a headroom diagnostic, explicit input and conditional comparisons, and causal tests with positive controls. The arithmetic results illustrate both the value and the limits of these checks. Probe-based truthfulness and deception monitors (Azaria and Mitchell, 2023; Orgad et al., 2025; Goldowsky-Dill et al., 2025) also motivate careful input comparisons, but predictive usefulness does not require causal participation in the behavior being monitored. Their practical reliability must be evaluated on relevant distributions and error costs.

2 Related Work

Accessibility, input reconstruction, and use. Information-theoretic and minimum-description-length treatments clarify what a probe measures (Pimentel et al., 2020; Voita and Titov, 2020). Decoding alone need not recover a known computation (Jonas and Kording, 2017); interpretability illusions and causal dissociations occur across modalities (Bolukbasi et al., 2021; Hanna et al., 2023; Makelov et al., 2024; Huang and Chang, 2025). Random-label controls test probe capacity (Hewitt and Liang, 2019). Our input control addresses a different ambiguity: whether the target is already predictable from specified prompt features. Conditional probing asks what a representation adds given those features (Xu et al., 2020; Hewitt et al., 2021); our wider-band reanalysis evaluates that comparison directly.

Concurrent work by Darade and Thorat (2026) finds candidate arithmetic intermediates more linearly accessible in activations than in raw-input baselines, including substantial accessibility at initialization. Their identified causal route does not transmit those candidates as the hypothesized staged algorithm predicts. This supports a route-specific distinction between accessibility and use, rather than establishing a definitively computed-but-unrouted intermediate. Our complementary contribution is the additive-baseline headroom calculation and the contrast between input prediction, answer-site transplants, and operand-site controls.

Interventions and arithmetic. Causal probing and concept erasure distinguish predictive from

behavioral evidence (Huang et al., 2023; Elazar et al., 2021; Belrose et al., 2023). Probes may exploit non-concept features (Kumar et al., 2022), and interventions trade completeness against selectivity (Canby et al., 2025). Activation-patching guidance therefore emphasizes controls and the limits of negative results (Zhang and Nanda, 2024; Heimersheim and Nanda, 2024).

Arithmetic work identifies structured number features (Wallace et al., 2019; Zhou et al., 2024; Kantamneni and Tegmark, 2025), operand-to-output pathways (Stolfo et al., 2023), heuristic computation (Nikankin et al., 2025), and causally effective arithmetic neurons (Du et al., 2025). In trained equation solvers, intermediate-direction interventions can change answers (Matsumoto et al., 2022). These results do not imply that every decodable arithmetic target has the same status at every site. Our behavioral controls also connect to prompt sensitivity (Sclar et al., 2024; Williamson et al., 2025; Sahoo et al., 2026). Few-shot recovery constrains the failure but does not, by itself, explain how demonstrations alter processing (Min et al., 2022).

3 Setup and Evaluation

Models and stimuli. We study base and instruction-tuned (“instruct”) Llama-3.1-8B (Grattafiori et al., 2024). Unless stated otherwise, prompts are bare arithmetic strings without a chat template. Greedy generation is scored by the first integer in the continuation; a parsing failure is incorrect. The primary battery uses 400 fixed-seed operand pairs from [20, 49]. C1 is $(0 + B) * C =$; comparison conditions are C4, $(B * C) =$, and C6, $(0 + B) =$. C3, $(B) * C =$, controls for parentheses without the added operation. Appendix H gives verbatim strings.

The study developed from an exploratory operator-precedence feasibility check comparing token-matched additive-identity expressions. The control battery exposed unequal difficulty between the candidate parses and motivated the present focus. The example was selected through that investigation, not sampled to estimate how often models fail at composition. Appendix O documents the retained discovery records and subsequent band and layer choices.

Probes and input baselines. The target is the scalar product BC . Regression R^2 measures squared-error reduction and is distinct from exact

integer recovery. Our original ridge probe uses training-fold centering and a fixed trace-scaled penalty. Its input baseline has two noisy copies of each scaled operand and 128 independent noise columns; no activations or interaction features enter it. The baseline asks what this declared additive feature class predicts, not which features the hidden-state probe actually uses.

For the wider-band reanalysis, we also use clean ordinary least squares on $[1, B, C]$. We compare this with hidden-state ridge and a joint linear probe on the clean operand features and hidden state. Five outer folds evaluate predictions; three inner folds select a C4 decoding layer using only outer-training pairs. Penalties and joint-block weights are then selected inside the outer-training set. All centering and block normalization use training statistics. We report seed-0 pooled out-of-fold scores, with two additional fold partitions as sensitivity checks (Appendix M). This replaces the designed bands’ earlier bootstrap summaries, whose repeated items could cross training and test folds.

Headroom. For baseline score R_F^2 , the maximum possible standalone or conditional improvement is $1 - R_F^2$. Under independent operands and additive prediction, Prop. 1 supplies a population value from the input distribution. The analytical and empirical ceilings are different quantities: in [20, 49] they are 0.0305 and approximately 0.0321, respectively. A low ceiling limits the possible improvement; it does not establish that a study can resolve it statistically.

Interventions and outcomes. At a specified answer-site residual-post hook, steering moves h along a fitted probe direction $\hat{w} = w/\|w\|$: $h \leftarrow h + k(c(P') - \hat{w} \cdot h)\hat{w}$. The coordinate $c(P')$ makes the fitted scalar prediction equal the donor product P' at $k = 1$; $k \in \{2, 4, 8\}$ overshoots it. This changes the probe’s prediction, not an independently verified algorithmic variable. Full-state swaps instead transplant the donor’s entire residual vector. Donors have a distinct product and aligned token positions. Key-value patches and attention-edge knockouts test complementary interfaces.

We score the intervention’s change in preference for the donor product over the target’s true product:

$$\Delta = \log \frac{p_I(P')}{p_I(P)} - \log \frac{p_0(P')}{p_0(P)}. \quad (1)$$

Candidate probabilities are teacher-forced token-sequence scores. Positive Δ favors the donor rela-

Auditing a computational interpretation

1. Input comparison. Declare available input features, estimator, evaluation split, and headroom; measure standalone and conditional gains.

2. Causal test. Intervene on the specified representation and measure a discriminative outcome at the same layer and distribution.

3. Positive control. Establish sensitivity with a comparable intervention and outcome at a behaviorally effective site. A successful control strengthens a negative result’s interpretation within its tested scope.

Figure 1: Three checks, with an operational recipe in Appendix G. Predictive usefulness and causal participation are separate evaluation goals.

	surface	base	instruct
C0	$B * C =$	0.838	0.848
C3	$(B) * C =$	0.495	0.755
C4	$(B * C) =$	0.890	0.908
C6	$(\emptyset + B) =$	1.000	1.000
C1	$(\emptyset + B) * C =$	0.508	0.265
A1	$(\emptyset + B) + C =$	0.993	0.995

Table 1: Exact-match accuracy on 400 primary-band pairs. Base C3/A1 are from the separate wrapper run on the same pairs; remaining entries are from the main battery. Full intervals and template provenance appear in Appendices A–B.

tive to the unpatched run; e^Δ is the corresponding odds multiplier. Full-product scoring distinguishes products that share their first token. Generation additionally measures donor/true-answer emissions. Appendix L states scoring, donor, and sample conventions; Appendix N maps each probe and intervention to its actual layer and representation.

Uncertainty. Battery accuracies have Wilson 95% intervals; paired behavioral and historical intervention means use the released bootstrap summaries where specified. Item-percentile spreads and ranges across fold partitions are descriptive, not confidence intervals. We omit the contaminated probe-bootstrap ranges from revised evidential comparisons. A zero donor count is reported with its generation denominator, not as proof that the event is impossible.

4 A Syntax-Sensitive Arithmetic Failure

Both models are accurate on C4 multiplication and perfect on the isolated C6 wrapper, but less accurate on C1 (Table 1). The C4–C1 difference is 0.64 for instruct, with paired interval $[0.60, 0.69]$. However, C3 shows that base accuracy is already 0.495 with parentheses alone, close to C1’s 0.508. Instruct

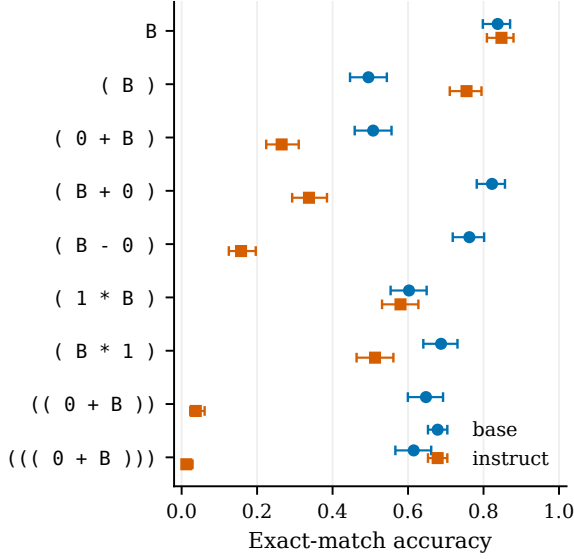


Figure 2: Multiplication accuracy for wrapper $\ast C =$ in the existing wrapper battery (400 pairs). Points and Wilson 95% intervals expose the parentheses control and left/right identity asymmetry. These are descriptive comparisons within each model.

instead falls from 0.755 on C3 to 0.265 on C1. The added operation therefore cannot be assigned the same interpretation in both models.

What the controls establish. Changing the outer operator to addition gives 0.993/0.995 accuracy (base/instruct). Equal token count alone therefore does not determine the failure, but this does not eliminate syntax or token-identity effects. Removing spaces worsens C1 markedly while unspaced C4 remains substantially more accurate. Extra nesting changes both multiplication and addition, with different effects across models. Within every matched product-size bin, instruct C1 remains below C4 (Appendix A). These observations constrain simple length and magnitude accounts without isolating a general composition deficit.

Wrapper placement matters: base accuracy is 0.823 for $(B + 0) \ast C =$, compared with 0.508 for the left-zero form; instruct is impaired by both (Figure 2). Instruction tuning is therefore neither uniformly beneficial nor uniformly harmful across forms.

Recovery and its limits. Four examples restore C1 accuracy to 0.883/0.915 in the original 400-item battery. A later causal battery with a fixed four-example prefix also recovers performance (§6). Demonstrations establish that the failure is recoverable in these settings; they do not distinguish

band	population R_F^2	empirical R_F^2	C1 gap (B/I)
[20, 49]	0.9695	0.9679	-0.0002/ - 0.0031
[2, 99]	0.8644	0.8648	+0.0572/ + 0.0482

Table 2: Historical fixed-regularization point estimates; B/I denotes base/instruct. Peak C1 layers are L6 and L7, respectively, for both models. These are not selection-separated estimates. The earlier bootstrap ranges are omitted. The stress band’s C4 accuracy (0.71/0.73) is below the chosen 0.80 floor.

changes in parsing, task recognition, routing, or later arithmetic computation.

5 Examining the Answer-Site Decode

High regression performance needs a comparator. In the primary band, the original fixed-regularization product probe at the C1 answer site scores $R^2 = 0.9677$ for base and 0.9648 for instruct. The input-only noisy-operand baseline scores 0.9679. These peak-layer estimates are descriptive because the same sample selected the layer. They show that high R^2 alone is weak evidence for a distinct product intermediate under this evaluation; they do not establish equal performance for every possible probe. For $(0 + 23) \ast 47 =$, 1081 is the regression target, not a verified exact probe prediction.

Wider bands reveal predictive gains and estimator sensitivity. The designed bands were chosen after the exploratory stress-band result using analytical headroom and a C4 point-accuracy floor of 0.80: [12, 69] for base and [10, 79] for instruct. Reanalysis uses the saved 400 unique pairs per model, with evaluation items excluded from layer and penalty selection. Table 3 shows that the clean and noisy operand baselines are similar, while activations add predictive value. At the original penalty multiplier, C1 gains over the clean baseline are 0.0277/0.0359. Tuning the multiplier inside training folds increases them to 0.0688/0.0817. The joint probe gains 0.0716/0.0872 over the input baseline.

Across three fold partitions, C1 standalone tuned gains range from 0.0675 to 0.0688 for base and 0.0790 to 0.0817 for instruct. Joint gains range from 0.0707 to 0.0728 and 0.0826 to 0.0872. These are split-sensitivity ranges, not confidence intervals. The original designed-band bootstrap means (0.036/0.042) are superseded, not evidence of significance.

Model/site	F	noisy	h_1	h_{tuned}	$[F; h]$
base C1	0.918	0.917	0.946	0.987	0.990
base C4	0.918	0.917	0.961	0.990	0.995
instruct C1	0.900	0.901	0.936	0.982	0.987
instruct C4	0.900	0.901	0.959	0.989	0.993

Table 3: Selection-separated wider-band R^2 (seed 0, 400 items/model). F : clean operand OLS; noisy: original-feature baseline; h_1 : fixed ridge multiplier 1; h_{tuned} : inner-selected multiplier; $[F; h]$: joint probe. Base uses [12, 69] and instruct [10, 79]. These scores are not directly comparable to the historical protocol in Table 2.

The improved wider-band scores qualify any general operand-echo account. Yet high regression accuracy still differs from exact recovery: on seed-0 C1, the tuned hidden-state probes have $R^2 = 0.987/0.982$, but rounded exact-match rates of only 1/400 and 0/400. Their mean absolute errors are 88/132 product units. Appendix M reports all estimator variants and error metrics. These are predictive results; the causal interventions below use the primary band and do not test the new wider-band probes.

Answer-site interventions in the primary band. We test a sufficient, transplantable product-store interpretation using successful C4 and failing C1. Steering changes the fitted product coordinate; full-state swaps transplant the whole answer-site vector. The dose experiment uses L23 for base and L4 for instruct, selected for the largest top-dose C4 effect on the same evaluated items. These exploratory curves are not selection-adjusted and use layers distinct from the C1 peak-probe layer. Late full-state swaps sample layers through L29. The full correspondence is explicit in Appendix N; a score selected elsewhere is not evidence about the exact intervention layer.

Steering produces no donor emissions in the 40 generated items per condition. On base C4, mean donor-versus-true preference increases with dose; instruct C4 is weaker and non-monotone, and C1 effects are small (Appendix P). Full-state swaps produce no donor emissions at the layers where generation was measured (60 items per cell), while late C4 swaps emit the true answer on 0.90–0.95 of that evaluated subset (Appendix I). These are observed rates, not a claim that patching preserves every item’s original behavior.

Positive controls establish sensitivity at operand sites. C1 single-operand residual swaps shift pref-

erence by +13 to +18 nats over layers 2–12 and produce donor answers on 48%–58% of base items and 19%–35% of instruct items. These use one-operand counterfactuals, whereas the answer-site sweep uses whole-problem donor permutations; cross-experiment effect ratios are descriptive. Appendix Q retains the detailed residual-site comparison.

The C4 key-value experiment provides a closer interface-matched comparison (Figure 3). Replacing answer-site keys and values across all layers yields +0.93/ + 0.20 nats and 0/60 donor emissions for base/instruct. Replacing the same interface at both operand spans yields +22.0/ + 23.8 nats and 54/60 and 56/60 donor emissions. Thus the measurement can register a large, behaviorally effective transplant.

Interpretation. The tests provide evidence against a sufficient, transplantable store in the examined answer-site channels. They leave original operand paths intact, so recomputation may counteract a transplant; a reader may reject a donor state; and full-state changes need not isolate every learned subspace (Heimersheim and Nanda, 2024; Geiger et al., 2024). The probe-aligned/orthogonal decomposition is a relative intervention-effect comparison, not an additive mediation analysis (Appendix P). A matched causal experiment on a corrected positive-gap band and its selected probe layer remains necessary before extending the causal conclusion there.

6 What the Mechanism Evidence Establishes

Operand access is necessary. Blocking attention from the answer site and later positions to the original operands across all layers reduces C4 accuracy from 0.88 to 0.00 for base and 0.89 to 0.01 for instruct. A stricter per-key version reproduces the collapse, while a closing-parenthesis placebo does not (Appendix J). Blocking attention to “=” also harms C0 and C4. This establishes dependence across those formats, but does not identify an exclusively formatting function for that position.

A strong necessary read occurs at layer 15. In the $n = 100$ single-layer scan, blocking operand edges at L15 reduces C4 accuracy from 0.85 to 0.01 and from 0.90 to 0.06, respectively: 99%/95% of the all-layer drop. L16 has smaller drops of 0.22/0.28; other layers change accuracy by at most

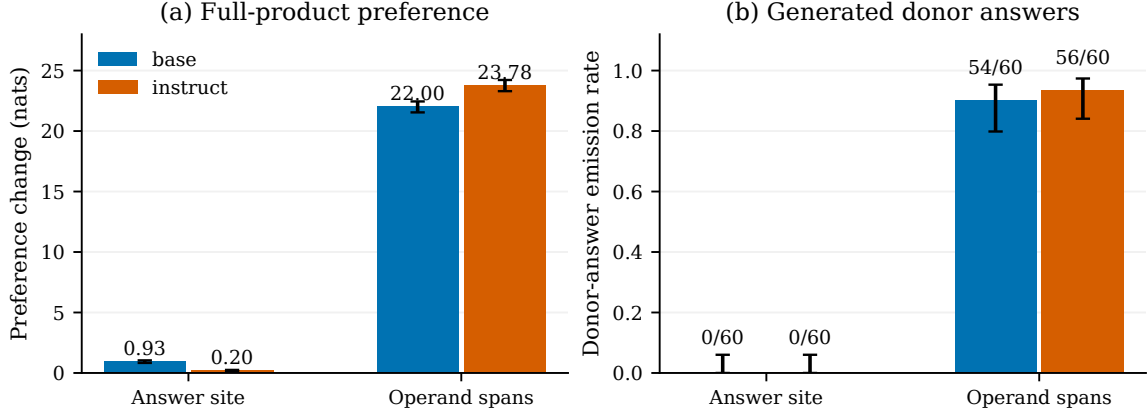


Figure 3: All-layer C4 key-value transplants on the primary band. **Left:** mean donor-versus-true log-odds change with the released 95% intervals ($n = 200$). **Right:** donor emissions among 60 generated items, with Wilson intervals. Answer-site and operand-span patches use the same interface and candidate-product metric. This control does not establish validity for every possible subspace intervention.

0.05. These are descriptive localization results over 32 comparisons, anchored by the all-layer control.

Replacing L15 operand value vectors yields $+8.5/+9.5$ nats, whereas key-only effects are near zero. Together with edge knockouts, this supports a content-bearing operand read. Query heads H14 and H15, sharing one key-value head, are jointly necessary for most of the drop. Their sufficiency has not been tested by retaining only their operand edges. Base and instruct share a fine-tuning lineage, so matching head identities are not independent mechanistic replications.

Operand probes already reach $R^2 \geq 0.94$ at the answer site by L13, yet removing access to the original operand positions at L15 collapses accuracy. The knockout localizes a necessary read of those positions; it does not locate multiplication itself. A weak transplant of L15’s summed attention output at “=” is consistent with later generated positions reading operands, but the combined-query knockout does not isolate their necessity. Appendix K reports the timing, head, and writer-search controls. A behavior-preserving test of the complete route remains open (Chan et al., 2022).

Few-shot evidence constrains recovery, not its cause. A fixed four-example prefix restores C1 to 0.900/0.887 on a separate 150-item causal battery. Removing target-operand access reduces both to zero. Answer-site swaps at L19, L23, and L27 give no donor emissions on the 60-item generation subset. Recovery therefore remains operand-dependent under these tests. Neither this result nor the successful-task L15 localization explains why zero-shot C1 initially fails, or which change

in processing enables recovery.

7 Scope and Practical Use

Behavioral transfer depends on format.

Among additional instruction-tuned models meeting the chosen C4/C6 point-accuracy floor of 0.80 in bare prompts, Gemma-2-9B-it scores C1 0.043 despite C4 0.99, and Llama-3.2-3B-Instruct scores C1 0.358. Three other models fail the floor and are reported separately. With native chat prompts, Gemma scores C1 0.95, while Llama-3.1-8B-Instruct falls below the C4 floor at 0.70 (Appendix E). This is behavioral evidence in specified formats, not cross-family transfer of the L15 mechanism.

On $(1 * B) * C =$, the original fixed-penalty analysis also gives near-zero standalone gaps ($-0.004/-0.011$) and small steering effects. Full-state swaps were not repeated for this wrapper. Its results broaden the tested surfaces but retain the estimator and intervention limitations above.

Applying the protocol beyond arithmetic.

For a truthfulness or deception monitor, first define the target, prediction time, and input information available to the probed site. Compare declared text baselines and activation probes on matched splits, including a joint predictor and relevant template/topic shifts. Report operational error rates as well as ranking quality. Causal interventions and positive controls are needed for a claim about computational participation; a useful predictor need not itself participate in producing the monitored behavior. The multiplication-specific additive formula does not transfer to AUROC, although the metric’s

upper limit still bounds improvement. Appendix G gives a concrete recipe. We have not run a realistic monitoring replication, so the arithmetic study does not establish how those applications will perform.

8 Conclusion

The same high regression score can have different evidential meanings depending on the input baseline, estimator, and evaluation distribution. In our narrow arithmetic band, the original probe is nearly matched by an input-only baseline; in wider bands, corrected standalone and joint probes add predictive value, particularly after regularization tuning. Neither result alone identifies computation or causal use.

Primary-band answer-site interventions have small average effects and no observed donor-answer emissions, while operand-site controls strongly affect behavior and operand access is necessary. This constrains the tested transplantable-store interpretation without resolving every causal route. Explicit baselines and headroom, representation-matched interventions, and positive controls make such conclusions more informative precisely when their scope and remaining gaps are stated.

Limitations

Our mechanistic evidence concerns Llama-3.1-8B base and instruct, which share a training lineage. The behavioral phenomenon depends on surface form and chat format; it does not isolate a general inability to compose operations. The example and later bands emerged through exploratory investigations. A point-accuracy floor does not guarantee population competence, and nested reanalysis of fixed datasets does not remove prior dataset-selection effects.

The corrected designed-band analysis separates layer and penalty selection from evaluation, but uses only 400 cached items per model and three fold partitions. Their spread is sensitivity to partition choice, not a formal uncertainty interval or independent replication. The primary and stress-band activation caches were not available for equivalent refitting; their retained fixed-penalty, peak-layer point estimates remain descriptive. Regularization materially affects the wider-band advantage, so the original narrow-band comparison should not be generalized to all estimators. Historical enriched-baseline tests cover specified feature constructions

rather than every finite-sample additive learner.

Causal experiments use the primary band, where the original probe has little standalone advantage. They do not test the corrected positive-gap bands or their newly fitted probes. Full-state sweeps stop at L29; L31 outputs are excluded because their cross-model duplication suggests a caching artifact. Other model/checkpoint metadata in historical captures are incomplete, as disclosed in Appendix L.

Positive controls establish sensitivity in their tested settings. They cannot exclude reader-side gating, compensatory recomputation, or a subspace effect obscured by wholesale transplantation (Geiger et al., 2021; Heimersheim and Nanda, 2024). Zero observed donor emissions are finite-sample observations. The L15 results localize necessary operand access, not multiplication, a complete sufficient circuit, or the cause of zero-shot C1 failure. The H14/H15 pair’s sufficiency and a position-restricted account of few-shot recovery remain untested.

A realistic truthfulness or deception replication would require task-specific input access, labels, generalization tests, and causal controls. We do not infer that published monitoring successes are operand echoes, or that causal participation is necessary for a monitor’s practical utility.

Availability

The original experimental code, result summaries, and figure scripts are available through the review artifact at <https://anonymous.4open.science/r/decodability-artifact-release-51FA/README.md>. The linked review artifact documents the historical experiments; the corrected reanalysis methods and results are described in Appendices L–M. Model weights are not redistributed. Historical automated verdict strings are retained as provenance and do not supersede the manuscript’s qualified interpretations.

References

- Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.

- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. [LEACE: Perfect linear concept erasure in closed form](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. 2021. [An interpretability illusion for BERT](#). *arXiv preprint arXiv:2104.07143*.
- Marc E. Canby, Adam Davies, Chirag Rastogi, and Julia Hockenmaier. 2025. [How reliable are causal probing interventions?](#) In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 857–878.
- Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. [Causal scrubbing: a method for rigorously testing interpretability hypotheses](#). AI Alignment Forum.
- Ishita Darade and Sushrut Thorat. 2026. [Represented is not computed: A causal test of candidate algorithmic intermediates in a transformer](#). *arXiv preprint arXiv:2605.22488*.
- Wenyu Du, Rui Zheng, Tongxu Luo, Stephen Chung, and Jie Fu. 2025. [Fine-grained manipulation of arithmetic neurons](#). In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 467–479, Suzhou, China. Association for Computational Linguistics.
- Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. [Amnesic probing: Behavioral explanation with amnesic counterfactuals](#). *Transactions of the Association for Computational Linguistics*, 9:160–175.
- Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. [Causal abstractions of neural networks](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. 2024. [Finding alignments between interpretable causal variables and distributed neural representations](#). In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pages 160–187. PMLR.
- Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. 2025. [Detecting strategic deception with linear probes](#). In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 19755–19786. PMLR.
- Leo A. Goodman. 1960. [On the exact variance of products](#). *Journal of the American Statistical Association*, 55(292):708–713.
- Aaron Grattafiori et al. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Michael Hanna, Roberto Zamparelli, and David Mareček. 2023. [The functional relevance of probed information: A case study](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 835–848, Dubrovnik, Croatia. Association for Computational Linguistics.
- Stefan Heimersheim and Neel Nanda. 2024. [How to use and interpret activation patching](#). *Preprint*, arXiv:2404.15255.
- John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. 2021. [Conditional probing: measuring usable information beyond a baseline](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639. Association for Computational Linguistics.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Wassily Hoeffding. 1948. [A class of statistics with asymptotically normal distribution](#). *The Annals of Mathematical Statistics*, 19(3):293–325.
- Jing Huang, Atticus Geiger, Karel D’Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. [Rigorously assessing natural language explanations of neurons](#). In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.
- Lianghuan Huang and Yingshan Chang. 2025. [Causality \$\neq\$ decodability, and vice versa: Lessons from interpreting counting ViTs](#). *Preprint*, arXiv:2510.09794.
- Eric Jonas and Konrad Paul Kording. 2017. [Could a neuroscientist understand a microprocessor?](#) *PLOS Computational Biology*, 13(1):e1005268.
- Subhash Kantamneni and Max Tegmark. 2025. [Language models use trigonometry to do addition](#). *Preprint*, arXiv:2502.00873.
- Abhinav Kumar, Chenhao Tan, and Amit Sharma. 2022. [Probing classifiers are unreliable for concept removal and detection](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

- Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. 2024. [Is this the subspace you are looking for? an interpretability illusion for subspace activation patching](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Yuta Matsumoto, Benjamin Heinzerling, Masashi Yoshikawa, and Kentaro Inui. 2022. [Tracing and manipulating intermediate values in neural math problem solvers](#). In *Proceedings of the 1st Workshop on Mathematical Natural Language Processing (MathNLP)*, pages 1–6, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. [Rethinking the role of demonstrations: What makes in-context learning work?](#) In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2025. [Arithmetic without algorithms: Language models solve math with a bag of heuristics](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szepkter, Hadas Kotek, and Yonatan Belinkov. 2025. [LLMs know more than they show: On the intrinsic representation of LLM hallucinations](#). In *The Thirteenth International Conference on Learning Representations (ICLR)*.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. [Information-theoretic probing for linguistic structure](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.
- Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. 2026. [Linear probes detect task format, not reasoning mode in language model hidden states](#). *arXiv preprint arXiv:2606.02907*.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. [Quantifying language models’ sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. [A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.
- Elena Voita and Ivan Titov. 2020. [Information-theoretic probing with minimum description length](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.
- Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. [Do NLP models know numbers? probing numeracy in embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.
- Dane A Williamson, Yangfeng Ji, and Matthew B. Dwyer. 2025. [Syntactic blind spots: How misalignment leads to LLMs’ mathematical errors](#). In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 1–14, Suzhou, China. Association for Computational Linguistics.
- Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. [A theory of usable information under computational constraints](#). In *The Eighth International Conference on Learning Representations (ICLR)*.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.
- Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. 2024. [Pre-trained large language models use Fourier features to compute addition](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Full Battery and Confidence Intervals

Table 4 gives the released template-matched battery with Wilson 95% intervals ($N = 400$). The released base battery omits C2, C5, and A2. Early feasibility notes contain base C2/C5 estimates, which are not combined with this release. Base A1, C3, and D1 are separate wrapper runs on the same pairs (0.9925, 0.495, 0.6475; Appendix B).

Magnitude control. Within matched $|B \cdot C|$ bins (instruct), C1 vs C4: [440, 813] 0.53/0.96 ($n=79$); [813, 1185] 0.28/0.94 (126); [1185, 1558] 0.16/0.88 (104); [1558, 1930] 0.19/0.89 (64); [1930, 2303] 0.00/0.70 (27). Every paired difference has a positive bootstrap CI.

Few-shot recovery. A fixed-seed 0/2/4-shot prefix restores C1 toward the C4 ceiling: base 0.508 $\rightarrow$ 0.838 $\rightarrow$ 0.883, instruct 0.265 $\rightarrow$ 0.885 $\rightarrow$ 0.915. This is recovery on the tested

	surface	base	instruct
C0	$B * C =$	0.838 [0.80,0.87]	0.848 [0.81,0.88]
C1	$((\emptyset + B) * C =$	0.508 [0.46,0.56]	0.265 [0.22,0.31]
C2	$((\emptyset + B * C) =$	—	0.678 [0.63,0.72]
C3	$(B) * C =$	—	0.755 [0.71,0.79]
C4	$(B * C) =$	0.890 [0.86,0.92]	0.908 [0.88,0.93]
C5	$\emptyset + B * C =$	—	0.542 [0.49,0.59]
C6	$(\emptyset + B) =$	1.000 [0.99,1.00]	1.000 [0.99,1.00]
C7	$(\emptyset + B) * C =$	0.018 [0.01,0.04]	0.190 [0.15,0.23]
C8	$(B * C) =$	0.765 [0.72,0.80]	0.785 [0.74,0.82]
A1	$((\emptyset + B) + C =$	—	0.995 [0.98,1.00]
A2	$\emptyset + (B + C) =$	—	0.928 [0.90,0.95]
D1	$((\emptyset + B)) * C =$	—	0.038 [0.02,0.06]

Table 4: Full battery, exact-match accuracy with Wilson 95% CIs. Note C5: relying on bare operator precedence ($\emptyset + B * C =$, 0.542) beats the parenthesized additive identity (C1, 0.265), so the parenthesization actively hurts (instruct).

wrapper surface, without identifying the mechanism of recovery.

Depth affects addition too. The doubly nested addition $((\emptyset + B)) + C =$ falls to 0.25 (instruct) from its single-nest form A1 (0.995), so an extra nesting is a general parsing stressor rather than further evidence about multiplicative composition.

B Wrapper Controls

Figure 2 maps multiplication accuracy by wrapper placement and family (wrapper_map_*.json). The principal wrapper comparison is now shown in the main text. The complete released map also contains right-multiplicative identity, subtraction by zero, and deeper nesting variants, with addition controls. The base and instruct map uses the same fixed operand pairs, but the separate base wrapper run is identified explicitly rather than silently combined with the main battery.

C By-Layer Decodability

Figure 4 traces operand decodability across all probed layers.

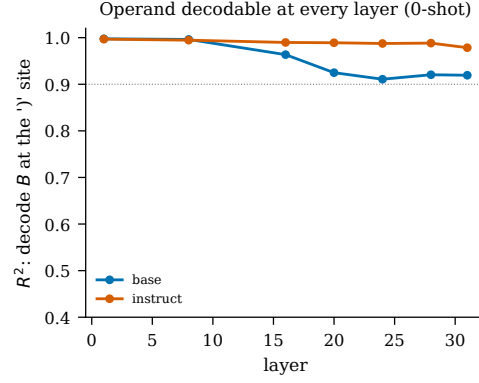


Figure 4: Operand B decodability from the C1 closing-parenthesis residual-post state in the failing zero-shot regime. Selected layers are plotted; the released full 32-layer curve stays above 0.91 base / 0.97 instruct.

D First-Token Steering Controls

Figure 5 collects the first-token steering nulls (causal_steering_*.json).

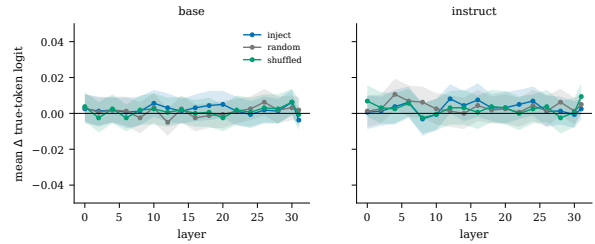


Figure 5: First-token steering at the C1 “=”: probe direction (inject), a norm-matched random direction (random), and a wrong product target (shuffled). Of the 102 plotted 95% intervals for mean ground-truth-token logit change, 100 include zero. The two exceptions are random/shuffled controls whose lower endpoints are below 0.01. Full-product effects and donor emissions are reported separately in Figure 7b.

E Cross-Model Battery

Each battery uses $N = 400$, and the additional models are instruction-tuned. In matched bare format, models are in scope if $C4$ and $C6 \geq 0.80$: Llama-3.1-8B base C1 0.508 / instruct 0.265; Gemma-2-9B-it 0.043 (C4 0.99); Llama-3.2-3B-Instruct 0.358 (C4 0.88). Out of scope (parts fail): Qwen2.5-7B-Instruct (C4 0.00), Mistral-7B-Instruct-v0.3 (C4 0.62), and Llama-3.2-1B-Instruct (C4 0.31). In chat format, Gemma-2-9B-it succeeds on C1 (C1 0.95, C4 0.99), but Llama-3.1-8B-Instruct and Llama-3.2-3B-Instruct fall out of scope (C4 0.70 and 0.74). No matched format contains both a qualifying C1-successful

model and a qualifying model with the C1 failure. Sources: `cross_model_battery_ci.csv` and `chat_format_control_ci.csv`.

F Headroom and the Additive Baseline

Headroom bound. Let R_{real}^2 be the activation probe’s score and R_F^2 the input baseline’s score for the same target. Since $R_{\text{real}}^2 \leq 1$,

$$R_{\text{real}}^2 - R_F^2 \leq 1 - R_F^2. \quad (2)$$

We call $1 - R_F^2$ the *headroom*. It also bounds the gain of a joint probe using hidden state h together with input features F , $R_{[h;F]}^2 - R_F^2$. The original gap compares separately fitted probes on the same folds; the synthetic baseline is specified in Appendix L. Proposition 1 supplies the population headroom for our additive-product setting before a model run. In population, best linear predictors over nested feature sets satisfy $R_{[h;F]}^2 \geq R_{\text{real}}^2$, so the separate-probe gap lower-bounds the conditional increment. Finite-sample cross-validated ridge can violate that ordering through estimation noise, so we report the empirical standalone and conditional increments separately in Table 3.

Proposition 1 (Additive baseline in closed form; classical). *Let the operands B and C be independent and non-degenerate with means μ_B, μ_C , variances σ_B^2, σ_C^2 , and finite second moments, and let the target be the product $B \cdot C$. Write $\tilde{B} = B - \mu_B$, $\tilde{C} = C - \mu_C$, and let $\mathcal{A}$ be the additive class $\{f(B) + g(C)\}$ with $\mathbb{E}[f(B)^2], \mathbb{E}[g(C)^2] < \infty$. Then (i) $\text{Var}(B \cdot C) = \sigma_B^2 \sigma_C^2 + \sigma_B^2 \mu_C^2 + \mu_B^2 \sigma_C^2$ (Goodman, 1960); (ii) the centered interaction $\tilde{B}\tilde{C}$ is orthogonal under $\langle X, Y \rangle = \mathbb{E}[XY]$ to every member of $\mathcal{A}$, constants included (the first-order Hoeffding decomposition; Hoeffding, 1948); and (iii) the best additive predictor of $B \cdot C$ is the linear one, $\mu_C B + \mu_B C - \mu_B \mu_C$, so the best-additive and best-linear baselines coincide at $R_{\mathcal{A}}^2 = 1 - \sigma_B^2 \sigma_C^2 / \text{Var}(B \cdot C)$. These results are classical; we repurpose them as a pre-probing design statistic.*

Proof of Proposition 1. Decompose $B \cdot C = \mu_B \mu_C + \mu_C \tilde{B} + \mu_B \tilde{C} + \tilde{B}\tilde{C}$. For any f with $\mathbb{E}[f(B)^2] < \infty$, independence gives $\mathbb{E}[\tilde{B}\tilde{C}f(B)] = \mathbb{E}[\tilde{B}f(B)] \mathbb{E}[\tilde{C}] = 0$ (the factor $\tilde{B}f(B)$ is integrable by Cauchy–Schwarz); symmetrically for $g(C)$, and constants are the case f constant, giving (ii). The first three terms of the decomposition form a linear predictor $\ell \in \mathcal{A}$

with residual $B \cdot C - \ell = \tilde{B}\tilde{C}$, so for any $s \in \mathcal{A}$, since $\ell - s \in \mathcal{A}$ is orthogonal to that residual by (ii), $\mathbb{E}[(B \cdot C - s)^2] = \mathbb{E}[(\ell - s)^2] + \mathbb{E}[\tilde{B}^2 \tilde{C}^2]$: the minimizer over $\mathcal{A}$ is $s = \ell$, which is linear. Independence factorizes the residual moment, $\mathbb{E}[\tilde{B}^2 \tilde{C}^2] = \sigma_B^2 \sigma_C^2$, and $\mathbb{E}[\tilde{B}\tilde{C}] = 0$, so the best additive predictor leaves residual variance $\sigma_B^2 \sigma_C^2$, giving (iii). The three centered terms of the decomposition are mutually orthogonal ($\tilde{B} \perp \tilde{C}$ by independence; $\tilde{B} \perp \tilde{B}\tilde{C}$ is (ii) with $f(B) = \tilde{B}$, symmetrically for $\tilde{C}$), so their variances add, giving (i). The argument uses independence, not mere zero correlation: for uncorrelated operands only $\mathbb{E}[\tilde{B}\tilde{C}] = 0$ survives, while the mixed moments $\mathbb{E}[\tilde{B}^2 \tilde{C}]$, $\mathbb{E}[\tilde{B}\tilde{C}^2]$, and $\mathbb{E}[\tilde{B}^2 \tilde{C}^2]$ need not vanish or factorize. $\square$

Instantiation and empirical checks. For independent operands uniform on integers $[l, h]$, $n = h - l + 1$ and $\sigma^2 = (n^2 - 1)/12$. The population additive baseline is 0.96949 at $[20, 49]$ and 0.86438 at $[2, 99]$, with corresponding analytical headroom 0.03051 and 0.13562. These values require no activations. The historical noisy-ridge scores (0.96788 and 0.86483) instead describe a finite sample, noise realization, and estimator.

The actual pair lists are deduplicated and reject single-digit-by-single-digit pairs, so they are not independent population draws. Small empirical operand correlations do not prove independence. Table 10 gives a corrected input-only sensitivity analysis on the exact hashed lists. Its finite-sample results assess those declared estimators, not every possible additive learner. Adding a product interaction is a deliberately target-reconstructing positive control, not evidence about the model’s internal mechanism.

G Applying the Three Checks

1. Define the target, prediction time, available input prefix, and desired generalization. A baseline must not see future continuation tokens unavailable to the probed site.
2. Declare the input feature class and training budget before examining evaluation results. A text-baseline ladder could include bag-of-words features and a frozen reader. Greater expressivity changes the question; a reader that can itself solve the task is not a neutral measure of which features the model uses.
3. Use shared outer splits for input, activation, and

concatenated probes. Select feature processing, layers, and penalties inside training data. Report standalone and conditional increments separately, with uncertainty or sensitivity labels that match the procedure.

4. Hold out relevant templates, topics, or scenarios. For binary monitors, report AUROC and operating-point false-positive/false-negative behavior. The score’s upper bound limits possible improvement, but the product-specific variance formula does not apply to AUROC.
5. For a computational claim, intervene on the same representation and distribution that support the predictive claim. Validate the perturbation and a task-specific positive control; report competence changes and failures as well as successes.

For example, a deception study could pair independently labeled scenarios with input-only transcript features, activations available at a stated prediction time, and their joint predictor, evaluating scenario-held-out performance. Adding concept-specific causal interventions would test the computational interpretation. A correlation-only pilot would implement the predictive checks alone. We have not run this experiment; monitoring utility can be established without showing causal participation in the monitored behavior.

H Verbatim Prompt Strings

Prompts use bare-continuation format with a beginning-of-sequence token (prepend_bos=True); below, $B=23$, $C=47$, and spacing is exact.

```
C0: 23 * 47 =
C1: ( 0 + 23 ) * 47 =
C2: ( 0 + 23 * 47 ) =
C3: ( 23 ) * 47 =
C4: ( 23 * 47 ) =
C5: 0 + 23 * 47 =
C6: ( 0 + 23 ) =
C7: (0+23)*47=
C8: (23*47)=
A1: ( 0 + 23 ) + 47 =
A2: 0 + ( 23 + 47 ) =
D1: (( 0 + 23 )) * 47 =
```

The few-shot C1 template prepends k solved C1 lines (distinct operands, correct answers), newline separated, then the target C1 line; e.g. $(0 + 20) * 21 = 420$ \n $(0 + 23) * 47 =$.

I The Answer Site Under Full-State Swaps

After the full-state swap at the C4 “=” site, no donor-product emissions are observed ($\text{emit-}P' = 0$ at every sampled odd layer from L19 through L29). The model emits the true answer on 0.90–0.95 of the evaluated subset across those layers (Table 5). The transplanted state is not sufficient to determine the emitted answer. We exclude L31 from the sweep on two grounds: a swap at the final layer of the last prompt position directly overwrites the next-token logit input rather than testing stored state, and the released L31 cells are bit-identical across the two models, which suggests a caching artifact in that run; no claim rests on them. High emit-true rates are consistent with later generated positions rereading operands, whose paths remain intact. The combined-query knockout does not isolate this explanation from other compensatory routes (Appendix K). The first generated token can be shared by two candidate products, motivating the full-sequence metric.

layer	base		instruct	
	Δ (nats)	emit-true	Δ (nats)	emit-true
19	+0.59	0.90	+0.12	0.95
23	+0.66	0.92	+0.15	0.92
27	+0.53	0.92	+0.13	0.93
29	+0.23	0.93	+0.05	0.93

Table 5: Full-state swap at C4 “=”: the $\text{emit-}P'$ column is omitted; it is 0 at every tested layer. L21 and L25 were also run and show the same pattern; the model keeps emitting its own correct answer. L31 is excluded (see text). `causal_steering_lateswap_*.json`.

J The Readout-Timing Test

Mechanics. Attention at layer l computes keys and values before that block’s residual-post hook. The first answer token is computed at the “=” position itself, and its path runs through the swapped L19 residual, but later generated tokens can also attend to cached “=” keys and values, including those computed at L19 before the hook. The key-value test writes a paired donor item’s keys and values (`hook_k/hook_v`; `hook_k` is captured before rotary positional encoding, and items are token-aligned so same-position writes share the rotary phase) at the C4 “=” index over a layer window, and scores the full-product metric of §3; the positive control writes the same interface at the operand-digit spans over all layers. At a residual-post hook

at L19, the L19 keys and values have already been computed. The early [0, 18] window therefore tests the layers strictly before the hook. The [19, 31] window directly replaces L19 and all later answer-site key-value channels; the all-layer window additionally includes [0, 18]. The all-layer patch also replaces what “=” reads from itself via self-attention. The knockout test zeroes the post-softmax attention pattern (no renormalization) at every layer for the stated query-key sets, with greedy exact match over $n = 200$ and paired-bootstrap drops. The operand-edge knockout drops accuracy by $+0.880$ [$+0.835, +0.925$] in both models; the “=”-edge drops are $+0.810$ [$+0.750, +0.865$] for base and $+0.890$ [$+0.845, +0.930$] for instruct. The C0 control shows that the “=”-edge damage occurs in both tested formats; this does not isolate an exclusively formatting function. Emit rates in Table 6 use a 60-item generation subsample whose unpatched accuracies, 0.82 for base and 0.88 for instruct, sit slightly below the 400-item battery values, consistent with subset noise under the Wilson intervals.

Intervention sensitivity and the instruct asymmetry. The interface-matched all-layer K/V comparison gives answer-site shifts of $+0.93/+0.20$ nats and operand-span shifts of $+22.0/+23.8$ for base/instruct (Table 6). The operand-to-answer mean-effect ratios are approximately 24 and 119, respectively. They show that this metric registers large effects at operand sites; they are descriptive ratios, not causal shares or a formal test of a between-model difference.

window	model	Δ [95% CI]	emit- P'	emit-true
= [0, 18]	base	$+0.02$ [$-0.00, +0.04$]	0	0.85
	instruct	$+0.01$ [$-0.01, +0.03$]	0	0.90
= [19, 31]	base	$+0.92$ [$+0.82, +1.02$]	0	0.82
	instruct	$+0.19$ [$+0.14, +0.24$]	0	0.88
= [0, 31]	base	$+0.93$ [$+0.83, +1.04$]	0	0.82
	instruct	$+0.20$ [$+0.15, +0.25$]	0	0.88
operands [0, 31]	base	$+22.0$ [$+21.5, +22.4$]	0.90	0
	instruct	$+23.8$ [$+23.3, +24.2$]	0.93	0

Table 6: Key-value-interface patch at C4 “=”, measured as the donor-versus-true log-odds shift (nats; $n = 200$, emissions on the 60-item subsample). The early window brackets zero; neither the late nor all-layer answer-site patch emits the donor product. The all-layer patch through the same interface at the operand spans moves $+22/+24$ nats and installs the donor answer. The late window tests 13 key-value layers, a different intervention from a single residual-post swap. kv_patch_equals_*.json.

condition	base acc [CI]	instruct acc [CI]
C4 baseline	0.880 [0.835, 0.925]	0.890 [0.845, 0.930]
block operands	0.000	0.010
block “=”	0.070	0.000
block “)” (placebo)	0.865	0.905
block operands (per-key)	0.000	0.000
C0 baseline	0.825 [0.770, 0.875]	0.840 [0.790, 0.890]
C0 block “=”	0.000	0.000

Table 7: Attention-edge blocking during generation (greedy exact match, $n = 200$; parse-fail 0.000 everywhere; paired-drop CIs in the released file). The per-key variant confines each operand to its own residual stream. Cutting operand access collapses C4; cutting the “=” edge collapses both C4 and bare C0, showing dependence in both formats, while the early [0, 18] key-value content patch has no detectable effect. attn_knockout_*.json.

K Surface, Band, Readout, and Regime Tests

We report four follow-ups on fixed operand pairs (dataset hash verified) at $n = 200$ (swaps), 400 (band batteries and captures), 100 (per-layer knockouts), and 150 (few-shot), with seed 1009 and emission rates on 60-item subsamples; a notebook in the released mirror reproduces them end to end.

C1-surface answer-site swap. We run the same non-self donor pairing and full-state swap as in the released late-layer sweep on the failing C1 surface at sampled odd layers L9–L29 and on C4 at sampled odd layers L9–L17 plus L19 and L23 (Table 8). The C1 site peaks at $+0.19$ (base) / $+0.11$ (instruct), below the C4 peaks ($+0.73/+0.14$ in Table 8), with emit- $P' = 0$ at every measured layer on both surfaces. Reruns with newly paired donors at C4 L19 and L23 ($+0.64/+0.10$; $+0.73/+0.14$) replicate the released sweep. The C4 mid-layer means are near zero through L13 and become larger at L15, the layer highlighted by the knockout below. This is a descriptive correspondence. Bracketed spreads in the released files are per-item percentile bands, not confidence intervals of the mean. We interpret the layer means, the peak contrast, and the emit rates.

layer	C1 surface		C4 surface	
	base	instruct	base	instruct
9	+0.03	-0.00	-0.00	+0.01
11	+0.02	+0.01	+0.00	+0.01
13	+0.02	+0.02	+0.02	+0.02
15	+0.03	+0.02	+0.22	+0.05
17	+0.06	+0.04	+0.45	+0.07
19	+0.09	+0.06	+0.64	+0.10
21	+0.16	+0.08		
23	+0.19	+0.11	+0.73	+0.14
25	+0.14	+0.09		
27	+0.12	+0.09		
29	+0.05	+0.02		

Table 8: Answer-site full-state swap, mean donor-versus-true log-odds shift (nats; $n = 200$ per cell; emissions were measured on 60 items at the three layers per surface with largest absolute mean shifts, with 0/60 donor emissions in each measured cell; per-item percentile bands are in the files). The failing C1 surface peaks below the successful C4 surface, and neither emits the donor answer (the largest cell, +0.73, is an odds factor of $e^{0.73} \approx 2.1$). `c1_answer_swap_*.json`.

Closed-form band design. The two candidate bands were specified using population headroom after the stress-band result was observed. Base failed the [10, 79] C4 point-accuracy floor (0.7675), then passed at [12, 69] (0.8125, Wilson interval [0.7713, 0.8477]); instruct passed at [10, 79] (0.8200, [0.7794, 0.8545]). Both selected bands have C6 accuracy 1.000. The rule concerns observed point accuracy; the intervals cross 0.80.

The original notebook selected a C4 decoding peak but omitted its numeric layer from the output. Recomputing that original full-sample selection from the saved activations gives L6 for both models. The corrected analysis instead selects within each outer-training split (Appendix M). The historical bootstrap means and interval midpoints are superseded by Table 3; no causal intervention has been added on either designed band.

Few-shot locus. A fixed, shared four-shot prefix (examples (45, 37), (44, 28), (43, 41), (40, 22); tokenization boundary verified per item) restores C1 to 0.900 (base) / 0.887 (instruct), replicating the previously reported recovery. Under recovery, the operand-edge knockout collapses accuracy to 0.000 on both models, and the “=” full-state swap installs nothing (emit- $P' = 0$ at L19, L23, and L27). `fewshot_locus_*.json`.

Inside the bottleneck. Per-head edge knockouts at L15 (all 32 heads, $n = 100$; the all-head operand-edge knockout reproduces Figure 6, seed 1109) show that H14 and H15 individually reduce accu-

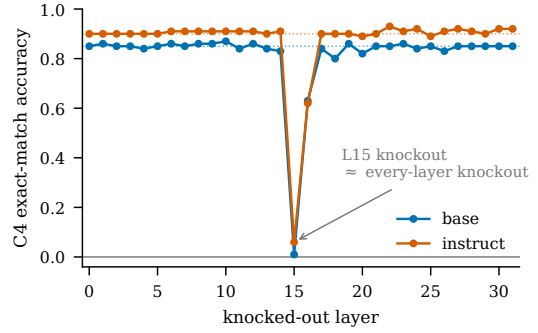


Figure 6: Layer-resolved operand-edge knockout on the C4 surface: accuracy when attention from “=” and all subsequent positions to the operand keys is zeroed at one layer at a time ($n = 100$; dotted lines are unablated baselines). Both models (which share a fine-tuning lineage) collapse at L15 alone, 99%/95% of the all-layer knockout’s effect, with a partial effect at layer 16 and every other layer near baseline. The other layers serve as within-curve controls. Descriptive curve (32 comparisons); the hypothesis test is the all-layer anchor (Table 7). `readout_layer_map_*.json`.

racy by 0.19/0.19 (base) and 0.11/0.14 (instruct). Knocking out both produces a 0.79 drop versus the 0.84 all-head drop for base, and 0.78 versus 0.84 for instruct. Cumulative sets of the top 4, 8, and 16 heads ranked by individual knockout drop give drops of 0.78, 0.79, and 0.78 for base, and 0.84, 0.84, and 0.85 for instruct. These set interventions are not additive estimates of individual head contributions. Joint sufficiency is untested: we have not spared the pair while cutting the other 30 heads’ operand edges at L15.

Each head has a partial effect alone, while their joint knockout produces a larger drop than the sum of the individual drops and nearly reproduces the all-head collapse. This is also consistent with the attribution-shortlisted single-head output patches finding no answer-writing locus; the same indices across models reflect their shared fine-tuning lineage, not independent replication, and with 32 heads tested at this n the claim rests on the cumulative knockout and survival through fine-tuning, not per-head significance. A donor value-patch at the operand positions at L15 alone moves +8.47/+9.55 nats (38–40% of the all-layer operand key–value effect through one layer’s values) while a key-patch control moves nothing (−0.003), so the L15 read carries operand *content*, not merely position selection (the value-patch spans all L15 heads; attributing the content read specifically to the pair combines separate interventions). Patch-

Claim	Assessment	Evidence and scope
Sufficient, transplantable product store at “=”	Evidence against in the tested answer-site channels	Full-state swaps have no observed donor emissions and retain the true answer on 0.90–0.95 of the evaluated subset; an all-layer answer-site key–value patch also has no donor emissions. Reader-side gating remains possible.
Answer assembled using operand information	Necessary operand read supported	Operand-edge knockout reduces C4 from 0.88 to 0.00 (base) and 0.89 to 0.01 (instruct), with the dominant necessary read at layer 15. This does not localize multiplication itself.
Sparse circuit writing the product at “=”	No answer-writing locus among shortlisted heads	Counterfactual single-head output patches over an attribution-selected shortlist of 8 heads per model, using the normalized-recovery metric defined in Figure 8, find no sufficient head (best recovery < 0.02). Because shortlisted heads were tested individually, the scan cannot exclude a multi-head writer, sparse or distributed; the necessary H14/H15 operand-read pair is not shown sufficient to write the answer.

Table 9: Claims, evidence, and scope. Every result traces to a released file (Appendices A–K).

ing the summed output of all L15 heads at “=” is weak (+0.21/+0.02, emit- P' = 0), which is consistent with operand reads at later generated positions. Separate answer-site-only and generated-position-only knockouts would be needed to isolate their necessity.

The “=” residual already decodes both operands at $R^2 \geq 0.94$ by L13, before the bottleneck, with no jump at L15: the knockout identifies a necessary access path despite earlier operand decodability. Zeroing each local multilayer perceptron’s output (mlp_out) at “=” for L15–L18 individually changes accuracy by at most 0.02: none of these four individual ablations produces the collapse. l15_circuit_*.json.

L Baseline Scope and Reproducibility

Input-only sensitivity. The identity $\text{gap}_{\text{linear}} - \text{gap}_{\text{enriched}} = R^2_{\text{enriched}} - R^2_{\text{linear}}$ allows baseline comparisons without activations. We recompute these comparisons on the original 400-pair lists using five outer folds, three inner folds, and the same penalty grid as the new activation analysis. Linear and interaction controls use OLS; other feature families use ridge with training-fold feature standardization. Quadratic adds B^2, C^2 ; Fourier adds sine/cosine terms at periods 2, 5, 10, 100; dense additive includes powers, roots, logarithms, inverses, and eight Fourier periods per operand. One-hot adds separate indicators for each possible operand value in the declared band. No family except the explicit interaction includes a joint operand feature.

These corrected out-of-fold comparisons replace the earlier enrichment bootstrap ranges and the endpoint-subtraction “gap bounds.” The historical files remain archived, including their heuristic ver-

Input estimator	[20, 49]	[2, 99]
Linear OLS	0.9685	0.8649
Quadratic ridge	0.9679	0.8643
Fourier ridge	0.9662	0.8606
Dense additive ridge	0.9654	0.8546
One-hot additive ridge	0.9607	0.7220
Interaction OLS	1.0000	1.0000

Table 10: Corrected input-only out-of-fold R^2 on the original pair lists (seed 0). The explicit interaction includes BC and is a target-reconstruction positive control. Appendix L specifies the feature families and evaluation procedure. No activation scores are refitted by this analysis.

dicts, but are not used as statistical certification. The saturated one-hot feature span can express nonlinear additive functions; finite-sample estimation can nevertheless perform worse than a smaller linear model. The corrected analysis also tests the known nonlinear additive target $B^2 + C^2$ as a positive control.

Historical environment and samples. Historical generation uses greedy decoding, first-integer parsing, and a beginning-of-sequence token, with base meta-llama/Llama-3.1-8B and the corresponding instruct model in bf16 on an A100. The release reports torch 2.11 and transformers 5.12, with TransformerLens 3.5.0 in the environment summary; some individual logs record its version as unknown. The instruct revision is recorded as 0e9e39f in that summary. Full model/tokenizer revisions are not recorded consistently for every capture, so the manifest marks missing values rather than assigning a common revision retrospectively.

The primary and stress lists have SHA-1 prefixes db4da1 and 3357af. Full-state swaps/component

interventions use $n = 200$ (generation on 60); doses use 200 (40); same-metric operand swaps use 160 (48); position-localization patches use 128; the attribution-selected eight-head search uses 48; the multiplication-by-one steering test uses 150. The key-value/knockout battery uses seed 909; later surface/readout/few-shot tests use 1009, and L15 head tests use 1109. The experiment table below distinguishes these datasets.

Scoring and donor conventions. Candidate products are scored as a leading space followed by the product’s token sequence, using each candidate’s own teacher-forced prefix, without an appended end-of-answer token. Thus these are sequence-prefix probabilities rather than probabilities of an exactly terminated integer response. Paired subtraction controls for each candidate pair’s unpatched score; it does not eliminate every possible length-dependent intervention effect. The inspected intervention helpers generate up to eight new tokens and parse the first integer; unparsable generations are incorrect. The later K/V and surface-swap protocols check distinct target products explicitly. The earlier complete-problem derangement guarantees distinct items; reconstructing its seed-707 split and seed-710 donor assignment on the released primary pair list also finds no equal-product collisions. Token positions are aligned within each tested interface. Answer-site sweeps derange complete problems; single-operand residual controls change one operand while holding the other fixed. Generation counts refer to the specified subsample, not all log-probability items.

For $0/n$ donor emissions, $3/n$ is an approximate one-sided 95% upper bound under the usual independent-sampling assumptions (about 0.05 for 60 and 0.075 for 40). Repeated layers on the same items are not additional independent observations; these are not simultaneous guarantees across the sweep. Per-item percentile bands describe effect heterogeneity, not uncertainty of a mean.

Original probe estimator and its replacement. The original probe centers features and targets on training folds and sets $\lambda = \text{tr}(K)/m$ with multiplier 1. Its synthetic baseline contains two columns each of $2(B - \bar{B}) + \epsilon$ and $2(C - \bar{C}) + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.5^2)$, plus 128 independent standard-normal columns. Subsequent training-fold centering removes the initial full-list mean offset. A fixed fold permutation is shared across compared estimators. Feature count and trace scaling do not

guarantee equal effective capacity.

The old item bootstrap reran CV after sampling rows with replacement, allowing identical activation items in training and test folds. It also re-generated noisy baseline features after resampling, treating duplicate activations and baseline noise asymmetrically. Relabeling these spans as descriptive does not repair their centers. We omit those spans and replace designed-band results with the selection-separated procedure below. The retained primary/stress fixed-fold point estimates are not contaminated by that bootstrap operation, but remain optimistic after peak-layer selection.

Provenance and exclusions. The historical source files are identified alongside the reported experiments. The corrected reanalysis is distinguished from historical estimates below. Exact cross-model duplication in the L31 late-swap rows motivates excluding those rows; no inference relies on them. Stimulus hashes of both designed-band caches match their result records, and both contain $400 \times 32 \times 4096$ answer-site tensors for C1 and C4 stored in float16. These checks establish stimulus/shape correspondence, not a complete reconstruction of missing checkpoint metadata.

M Selection-Separated Probe Reanalysis

We use the existing designed-band items, with no new model execution. Five outer folds evaluate predictions. Within each outer-training set, three inner folds select the C4 layer maximizing product R^2 among all 32 layers at the original penalty multiplier 1. The same selected layer is used for C1 and C4 comparisons. Each surface’s hidden probe then selects its ridge multiplier from $\{0.01, 0.1, 1, 10, 100\}$ using only those training items. The fixed-multiplier variant holds it at 1.

The clean baseline is OLS on $[1, B, C]$. A separate noisy baseline preserves the original construction with noise seed 920 fixed by original item and multiplier 1. For the concatenated clean-input and hidden-state probe, each centered block is normalized to unit mean squared norm using training statistics. The hidden block’s kernel weight is selected from $\{0.01, 0.1, 1, 10, 100\}$ together with the ridge multiplier on inner folds. This is a linear concatenated-feature probe with explicit block scaling, not a nonlinear input model. All choices are refitted in each outer fold; held-out targets enter neither selection nor fitting.

The primary partition has seed 0; seeds 17 and

42 repeat the entire procedure. Scores pool all 400 out-of-fold predictions. The ranges below vary partitions of the same dataset and are not confidence intervals. Recovered historical full-sample C4 peaks are L6/L6. Seed-0 outer-fold layers are [5, 6, 6, 5, 6] for base and [6, 6, 6, 6, 6] for instruct. Across all partitions, base uses L4–L6 and instruct uses L6. Hyperparameter selections, training/test indices, and predictions were recorded for verification.

Rounded exact match means rounding the real-valued probe prediction to the nearest integer and comparing it with *BC*. It differs from exact match on the model’s generated answer. The contrast between high R^2 and low rounded exact match is a property of these regression estimators and sample, not a universal limit on value decoding.

Implementation checks. Kernel predictions were checked against explicit primal ridge solutions, including joint block normalization. Changing held-out target values leaves fitted test predictions unchanged. Each original pair is unique; outer test folds are disjoint from all corresponding training/selection folds. CPU analysis uses NumPy 2.2.6 and SciPy 1.13.1. The grids and partition seeds are specified above. This reanalysis does not independently reproduce the original GPU runs.

N Experiment Correspondence

All layers are zero-indexed. “Answer site” means the last prompt token “=”; residual probes and swaps use the block’s residual-post tensor. K/V patches use the key/value hooks, a distinct interface. Attention-edge knockouts act after softmax without renormalization. The first output token is predicted at the answer site; subsequent output tokens have their own query positions and can still read original operands.

O Discovery and Selection Records

The retained feasibility notebook and run notes begin with a token-matched contrast between $(\emptyset + B) * C =$ and $(\emptyset + B * C) =$. The additive identity was intended to preserve the answer while changing operator depth. Behavioral checks revealed that it was not a reliable no-op for the model. The subsequent C0–C8 battery examined parentheses, placement, spacing, and the isolated wrapper; addition and nesting controls and an instruct-model run followed. The difficulty imbalance prevented

a clean precedence-localization interpretation and motivated the probing study.

This is an exploratory discovery account supported by the saved notebooks, README, and run records. They are not an exhaustive log of every informal prompt tried, and we do not claim preregistration. Early notes sometimes use stronger labels such as “composition-specific” or “computed”; the final interpretation is based on the reported controls. Tables 1 and 4 distinguish template-matched batteries from separate wrapper runs. All additional model configurations evaluated in the cross-model battery, including failures of the competence floor, are identified in Appendix E. Designed-band selection followed the stress-band observation and is described above; nested reanalysis changes estimator evaluation, not that discovery history.

P Component and Dose Controls

Dose probes were fitted on 200 training items. The candidate layers for both models were L4, L19, L23, L26, and L31. The layer maximizing C4’s mean effect at dose 8 on the 200 evaluation items was then used for the displayed C4 curve and C1 evaluation. C1 has its own fitted product direction at that layer. The released bootstrap intervals condition on the fitted direction and donor assignments; they are neither simultaneous nor adjusted for this layer selection. The component experiment’s L19 follows the historical late-swap selection rule; retained donor-emission rates tie at zero, so this is not evidence of a uniquely optimal layer.

The base full, orthogonal-only, and aligned-only mean shifts are 0.59269, 0.50214, and 0.08089 nats. Thus the orthogonal intervention retains about 85% of the full intervention’s mean shift. These are separate perturbations through a nonlinear downstream model, so the ratios do not identify additive causal shares. The instruct aligned interval is $[-0.00247, 0.02014]$, which is displayed in full. No component-only condition emits the donor product. A paired interaction-residual analysis would be needed for a stronger decomposition claim.

Q Additional Residual-Position Controls

The residual-site comparison retains the original effect measurements and localization controls. The answer-site example uses C4 while the single-operand controls use C1 and different donor construction. The all-layer C4 key–value experiment in

Model	Site	Estimator	R^2 [split range]	MAE	RMSE	Rounded exact
base	C1	Clean operand OLS	0.9184 [0.9170, 0.9184]	219.0	288.9	2/400
base	C1	Noisy operand, fixed	0.9175 [0.9158, 0.9175]	220.9	290.5	2/400
base	C1	Hidden, fixed	0.9460 [0.9344, 0.9460]	177.2	234.9	2/400
base	C1	Hidden, tuned	0.9871 [0.9856, 0.9871]	88.1	114.6	1/400
base	C1	Joint, tuned	0.9899 [0.9888, 0.9899]	79.8	101.5	0/400
base	C4	Clean operand OLS	0.9184 [0.9170, 0.9184]	219.0	288.9	2/400
base	C4	Noisy operand, fixed	0.9175 [0.9158, 0.9175]	220.9	290.5	2/400
base	C4	Hidden, fixed	0.9606 [0.9602, 0.9609]	150.6	200.6	3/400
base	C4	Hidden, tuned	0.9902 [0.9902, 0.9919]	72.1	99.9	1/400
base	C4	Joint, tuned	0.9953 [0.9947, 0.9953]	52.0	69.5	6/400
instruct	C1	Clean operand OLS	0.9002 [0.9002, 0.9039]	322.7	427.7	0/400
instruct	C1	Noisy operand, fixed	0.9008 [0.9008, 0.9047]	322.3	426.3	2/400
instruct	C1	Hidden, fixed	0.9361 [0.9343, 0.9386]	260.5	342.2	0/400
instruct	C1	Hidden, tuned	0.9819 [0.9819, 0.9841]	132.1	182.1	0/400
instruct	C1	Joint, tuned	0.9874 [0.9865, 0.9880]	114.1	152.3	1/400
instruct	C4	Clean operand OLS	0.9002 [0.9002, 0.9039]	322.7	427.7	0/400
instruct	C4	Noisy operand, fixed	0.9008 [0.9008, 0.9047]	322.3	426.3	2/400
instruct	C4	Hidden, fixed	0.9593 [0.9591, 0.9606]	211.0	272.9	1/400
instruct	C4	Hidden, tuned	0.9890 [0.9890, 0.9893]	107.0	141.8	1/400
instruct	C4	Joint, tuned	0.9929 [0.9928, 0.9930]	87.6	113.8	2/400

Table 11: Complete wider-band reanalysis. Point metrics use partition seed 0; brackets give the minimum and maximum R^2 over seeds 0, 17, 42, not a confidence interval. MAE/RMSE are in product units; exact recovery refers to the regression prediction, not model generation.

Experiment	Band / surface	Representation and layer	Selection, sample, and outcome
Original product probes	[20, 49] C1	Answer-site residual, L6 both	Peak over same five-fold CV scores; 400 items; descriptive R^2 .
Stress product probes	[2, 99] C1	Answer-site residual, L7 both	Same-sample peak selection; 400; descriptive R^2 .
Corrected designed bands	base [12, 69]; instruct [10, 79]; C1/C4	Answer-site residual, layer selected on training C4	Five outer / three inner folds; 400 unique pairs; standalone/joint regression and error metrics.
Dose steering	[20, 49] C1/C4	Answer-site residual, base L23; instruct L4	Largest C4 top-dose effect on the evaluated sample; separate surface probes; 200 scored, 40 generated; doses 1, 2, 4, 8.
Component transplants	[20, 49] C4	L19 answer-site residual; full, aligned, orthogonal	Original C4 probe direction; historical late-swap layer choice; 200 scored, 60 generated.
Late residual swaps	[20, 49] C4	Answer site, odd L19–L29	Whole-state donor transplant, no fitted probe; 200/60; L31 excluded.
Surface swap extension	[20, 49] C1/C4	C1 odd L9–L29; C4 odd L9–L17, L19/L23	200 scored; 60 generated at three layers selected by absolute mean shift per surface/model.
Single-operand control	[20, 49] C1	B or C residual, layers 2–12	One-operand counterfactuals; 160 scored, 48 generated; same units, different donor construction.
K/V control	[20, 49] C4	Answer site or both operand spans; all L0–L31	Same interface and donor metric; 200 scored, 60 generated; early/late windows also tested.
Operand-edge knockouts	[20, 49] C4	Queries at “=” and later, operand keys	All-layer 200; single-layer/head 100; L15 pair necessity, not sufficiency.
Four-shot causal battery	[20, 49] C1	Operand-edge KO; answer-site residual swaps L19/23/27	Fixed demonstrations; 150 behavioral items, 60 generation subsample for swaps.

Table 12: Correspondence between predictive and causal evidence. A decoded target at one layer/band is not treated as causally tested by an intervention at another. The wider-band fitted probes have no corresponding new causal test.

Figure 3 provides the interface-matched main-text comparison. Neither control universally validates every answer-site subspace intervention.

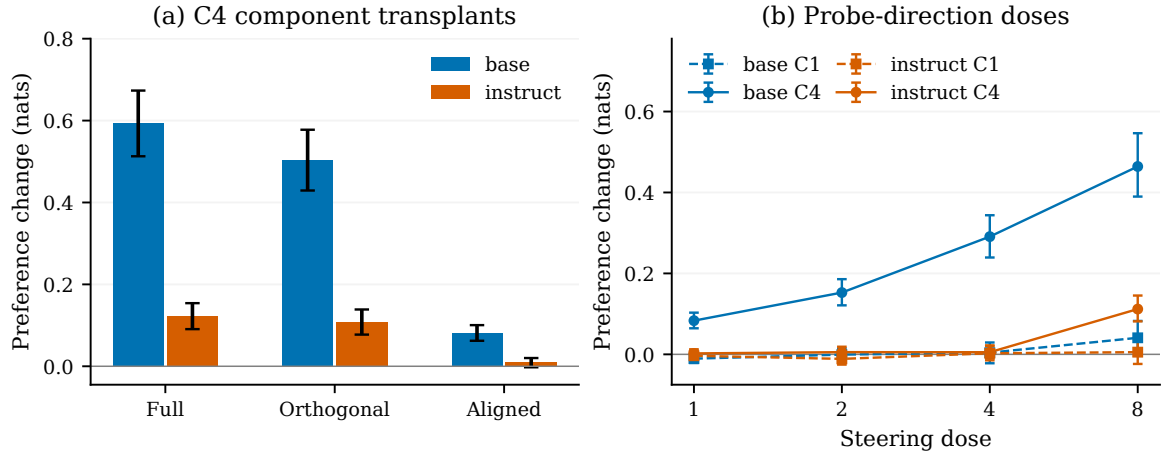


Figure 7: Historical primary-band controls with repaired layout. **(a)** C4 L19 component transplants; intervals are the released 95% summaries. The axis includes the instruct aligned component’s negative interval endpoint. **(b)** All four model/surface dose curves with their released intervals, using color for model and marker/line style for surface. Dose layers (base L23; instruct L4) were selected on the displayed C4 effects; intervals are pointwise and not selection-adjusted. Each condition has zero donor emissions on its generation subsample (60 for components, 40 for doses).

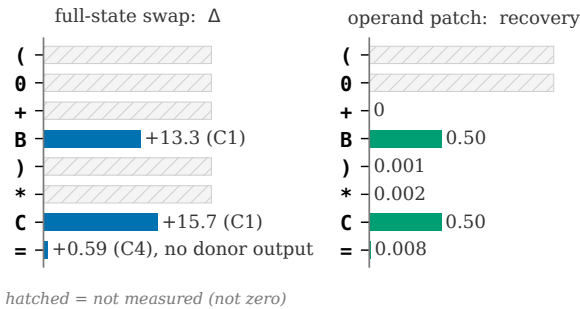


Figure 8: Why answer-site decodability does not establish computation (base model). **Left**, full-state swaps change donor-versus-true log odds by only +0.59 nats at C4’s “=”, where the model answers correctly, but by +13.3 and +15.7 at C1’s *B* and *C* positions; the answer-site swap never emits the donor product. **Right**, on C1, we replace one operand in a token-aligned counterfactual and patch in states from the original run. Normalized recovery is 0 for the counterfactual contrast and 1 for the original: about 0.50 at operand positions and about 0 at measured non-operand positions. This metric uses the full-product log-odds contrast; hatched cells are unmeasured. The tested operand positions have much larger effects; the cross-experiment scale comparison is descriptive.